HW1_STAT5000

Xingyu Chen

September 03, 2021

Due on Canvas on Friday, September 3, 2021 by 11:59 PM. Covers ex- ploratory data analysis. Instructions for "theoretical" questions: Answer all of the following questions. The theoretical problems should be neatly numbered, written/typed out, and solved. Remember to self-grade each question in a location, font, and color that is obvious, as well as including a total self-grade for the entire homework.

Total Score: 48/52

1 Theoretical Questions

1. Suppose that you would like to compare the Twitter habits of college students at CU-Boulder campus with the Twitter habits of college students more generally.

(a) What are the populations you are concerned with? Be specific! Q1: 2/2

Answer:

The college students in the CU Boulder who have Twitter habits and other college students who have Twitter habits.

(b) What is the relationship between these populations? Q2: 2/2

Answer:

They are all belong to the population of college students. They definitely have common features as an college students.

(c) What are some of the habits that you might consider? Pick three as an example. Q3: 2/2

Answer:

Daily News from campus; Local meetup events; Announcement from campus.

(d) If you had infinite time and resources, would you be able to measure these characteristics for every member of these populations? Q4: 2/2

Answer:

No, there are infinite features we can get from the data and some data may not visible to us due to privacy.

(e) Suppose that you don't have infinite time and resources; how would you go about estimating those population characteristics? Q5: 2/2

Answer:
Get sample from the population. Analyze the sample features so that I can get approximate conclusion for whole population.

2. 1000 American voters are contacted by telephone between the hours of 6pm and 8pm. The telephone numbers were chosen in the following way: pollsters put all of the numbers they had in their database in numerical order (e.g., the num- ber (555) 555-5555 was ordered before the number (555) 555-5556); then, they called every 100th number. Each person was asked whether he or she would vote for a Congressional candidate who supports universal health care. The point of this survey was to determine the level of support, amongst American voters, for universal health care.

(a) Identify the population. Q6: 2/2

Answer:
American voters for universal health care.

(b) Identify the sample. Q7: 2/2

Answer:
All the phone numbers in the data, each phone number represent one American voter.

(c) Identify the sample frame. Q8: 2/2

Answer:
Every 100th number of American voter in numerical order in the database.

(d) Identify the type of sample. Q9: 2/2

Answer:
Systematic sampling because they arrange them in numerical order then call every 100th.

(e) Identify the variable of interest. Q10: 2/2

Answer:
The vote for a Congressional candidate who support universal health care.

3. Let x1,. . . ,xn be a variable measured for units in a sample. Let μ denote the mean of the population from which the sample came. The mean of the sample, as always, is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$

(a) For what value of c is the quantity $\sum_{i=1}^{n}(x_i - c)^2$ minimized? Q11: 1/2

Answer:
if c = $\overline{x}$,
then $(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 = S^2$ Thus, when c = mean of the sample, the $\sum_{i=1}^{n}(x_i - c)^2$ minimized.

(b) Using the result from part (a), which of the two quantities $\sum_{i=1}^{n}(x_i - \overline{x})^2$ and $\sum_{i=1}^{n}(x_i - \mu)^2$ will be smaller than the other (assuming that $\overline{x}$!= μ)? Q12: 2/2

Answer:

For sample, the

$$S^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$$

. Meanwhile for population: the

$$S^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n)$$

.

Thus

$$S_s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1) > S_p^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n)$$

Thus, $\sum_{i=1}^{n}(x_i - \mu)^2$ is smaller than the other.

(c) Let $y_i = x_i - \overline{x}$, for i = 1, . . . , n. How do the values of

$$s^2$$

and s for the y_i's compare to the corresponding values for the x_i's? Prove your result. Q13: 2/2

Answer:

$y_i = (1/n)\sum_{i=1}^{n}(x_i - \overline{x}) = ((x_1 - \overline{x}) + (x_2 - \overline{x}) + ... + (x_n - \overline{x}))/n = 0$ Thus, $S_y = \sqrt{(\sum_{i=1}^{n}(y_i - \overline{y})^2)/(n-1)} = \sqrt{(\sum_{i=1}^{n}(y_i - 0)^2)/(n-1)} = \sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2)/(n-1)} = S_x$ $S_y^2 = (\sum_{i=1}^{n}(y_i - \overline{y})^2)/(n-1) = (\sum_{i=1}^{n}(y_i - 0)^2)/(n-1) = (\sum_{i=1}^{n}(x_i - \overline{x})^2)/(n-1) = S_x^2$. Thus, the value is equal.

(d) Let $z_i = (x_i - \overline{x})/s$ for i = 1, . . . , n. What are the values of the sample variance and sample standard deviation for the z_i's? Prove your result. Q14: 1/2

Answer:

$z_i = (x_i - \overline{x})/s = (x_i - \overline{x})/\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2)/(n-1)}$,

$\overline{z} = (1/n)\sum_{i=1}^{n}((x_i - \overline{x})/s)$, $S_z = \sqrt{(\sum_{i=1}^{n}(z_i - \overline{z})^2)/(n-1)} = 1$, $(S_z)^2 = (\sum_{i=1}^{n}(z_i - \overline{z})^2)/(n-1) = 1$.

2 Computational Questions

Instructions for "computational" questions: Your work should be neatly done and in- clude all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do not put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. All computations should be done using R, which can be downloaded for free at https://cran.r-project.org/. This is your first opportunity to get familiar with R, so please take your time on the problems that require it. **Also get familiar with R Studio and R Markdown, as both should be used and produce an easy way to code and output everything in a nice format.**

All of these questions can be solved in multiple ways using R, one of which is Tidyverse. Challenge yourself to learn the foundation for Tidyverse now. This homework section can be solved using our Day 2 Tidyverse lesson!

1. Access the babies data from https://www.stat.berkeley.edu/~statlabs/data/ babies.data and store the data frame as babies. A description of the variables can be found at help at https://www.stat.berkeley.edu/~statlabs/labs.html.

```
library(tidyverse)
babies <- read_table('babies.data')
```

(a) Create a "clean" dataset that removes subjects if any observations on the subject are unknown. Store this cleaned dataset as clean. HINT: Are there numerical values being used to designate unknown values? Q15: 2/2

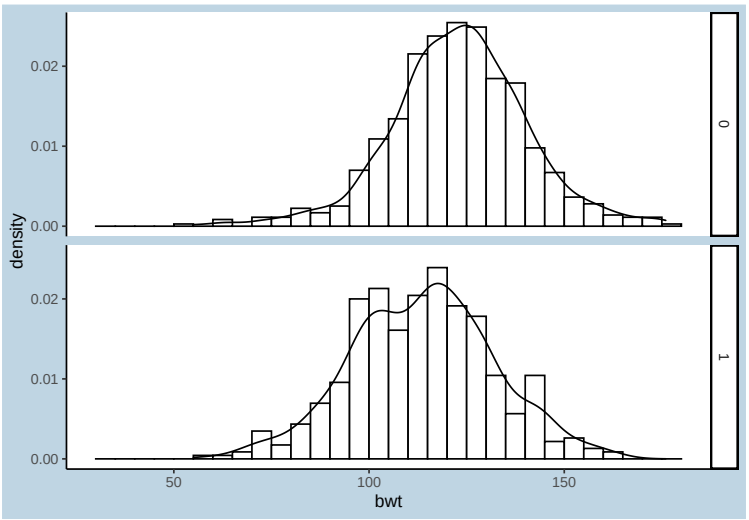
Answer:

```
clean <- babies[babies$bwt != 999 & babies$gestation != 999
                & babies$parity != 9 & babies$height != 99
                & babies$weight != 999 & babies$smoke != 9, ]
```

(b) Use the information in clean to create a density histogram of the birth weights of babies whose mothers have never smoked and another histogram placed directly below the first in the same graphics device for the birth weights of babies whose mothers currently smoke. Make the range of the x-axis 30-180 (ounces) for both histograms. Superimpose a density curve over each histogram. Q16: 2/2

Answer:

```
ggplot(data = clean, aes(x = bwt)) +
  geom_histogram(aes(y = ..density..),
                 breaks = seq(30, 180, by = 5),
                 colour = "black", fill = "white") +
  geom_density() +
  facet_grid(rows = vars(smoke)) +
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```



(c) Based on the histograms in (b), characterize the distribution of baby birth weight for

both non-smoking and smoking mothers. Q17: 1/2

Answer:

Non-smoking: The distribution of birth weights for the nonsmokers is negative skew, most of the babies have a birth weight between 120 to 130 ounces. It is distribute evenly.

Smoking: The distribution of birth weights for the nonsmokers is symmetric, most of the babies have a birth weight between 115 to 125 ounces. And it is not distribute evenly.

(d) What is the mean weight difference between babies of smokers and non- smokers? Can you think of any reasons not to use the mean as a measure of center to compare birth weights in this problem? Q18: 2/2

Answer:

```
mean(data.matrix(clean[clean$smoke == 0, "bwt"])) -
  mean(data.matrix(clean[clean$smoke == 1, "bwt"]))
```

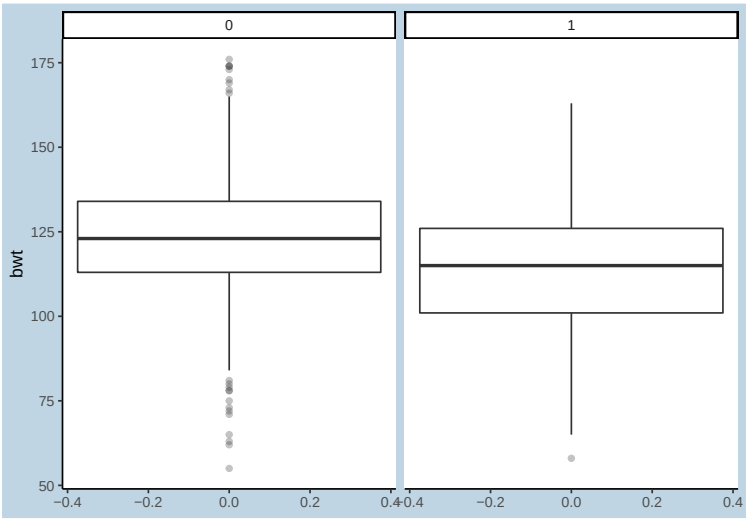
```
## [1] 9.261402
```

Reason not to use the mean: the birth weights is not distribute evenly for the smoking mother, so the mean of the population is not accurate reflect the features.

(e) Create side-by-side boxplots to compare the birth weights of babies whose mothers never smoked and those who currently smoke. Q19: 2/2

Answer:

```
ggplot(data = clean, aes(y=bwt)) +
  geom_boxplot(alpha=0.3) +
  facet_grid(~smoke) +
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```



(f) What is the median weight difference between babies who are firstborn and those who are not? Q20: 2/2

Answer:

```
median(data.matrix(clean[clean$parity == 0, "bwt"])) -
median(data.matrix(clean[clean$parity != 0, "bwt"]))
```

```
## [1] 2
```

(g) Compute the body mass index (BMI) for each mother in clean. BMI is defined as kg/m2 (0.0254 m = 1 in, and 0.45359 kg = 1 lb). Add the variables weight in kg, height in m, and BMI to clean and store the result in cleannp. Q21: 2/2

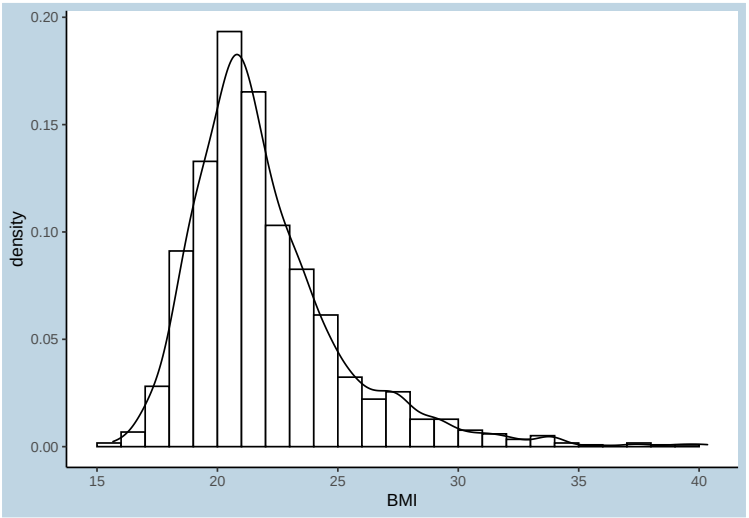
Answer:

```
clean$'weight in kg' <- clean$weight * 0.45359
clean$'height in m' <- clean$height * 0.0254
clean$BMI <- clean$'weight in kg' / (clean$'height in m' * clean$'height in m')
cleannp <- clean
```

(h) Characterize the distribution of BMI (e.g., symmetric, skewed). Q22: 2/2

Answer:

```
ggplot(data = cleannp, aes(x = BMI)) +
  geom_histogram(aes(y = ..density..),
                 breaks = seq(15, 40, by = 1),
                 colour = "black", fill = "white") +
  geom_density() +
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```



It is positive skew for the BMI distribution for each mother, most of population distribute between 20 to 21.

(i) Group pregnant mothers according to their BMI quartile. Find the mean and standard deviation for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke. Find the median and interquartile range for baby birth

weights in each quartile for mothers who have never smoked and those who currently smoke. Based on your answers, would you characterize birth weight in each group as relatively symmetric or skewed? Q23: 2/2

Answer:

```
cleannp_i <- cleannp
cleannp_i$quartile <- as.numeric(cut_number(cleannp_i$BMI, n = 5))

cleannp_nonsmoke <- filter(cleannp_i, smoke == 0)
cleannp_smoke <- filter(cleannp_i, smoke == 1)

cleannp_smoke_result <- cleannp_smoke %>%
  group_by(quartile) %>%
  summarize(mean = mean(bwt), sd = sd(bwt), min = min(bwt), max = max(bwt))

cleannp_nonsmoke_result <- cleannp_nonsmoke %>%
  group_by(quartile) %>%
  summarize(mean = mean(bwt), sd = sd(bwt), min = min(bwt), max = max(bwt))

cleannp_smoke_result
```

quartile	mean	sd	min	max
1	109.8136	19.13512	71	160
2	115.8242	16.87443	65	156
3	114.8675	18.22674	69	163
4	116.8824	17.13074	71	154
5	113.1566	19.04967	58	161

```
cleannp_nonsmoke_result
```

quartile	mean	sd	min	max
1	120.7607	15.00784	84	173
2	123.4615	18.95202	62	167
3	124.3901	15.19529	85	166
4	122.7248	17.33371	55	170
5	123.6316	19.46424	72	176

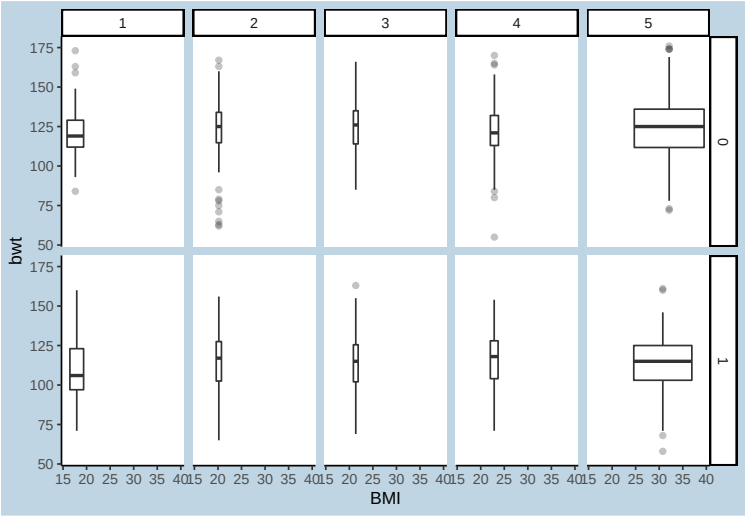
Each group is relatively symmetric.

(j) Create side-by-side boxplots of bwt based on whether the mother smokes conditioned on BMI quartiles. Does this graph verify your findings in the previous part? Q24: 2/2

Answer:

```
ggplot(data = cleannp_i, aes(x = BMI, y = bwt)) +
  geom_boxplot(alpha=0.3) +
  facet_grid(rows = vars(smoke),
             cols = vars(quartile)) +
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```

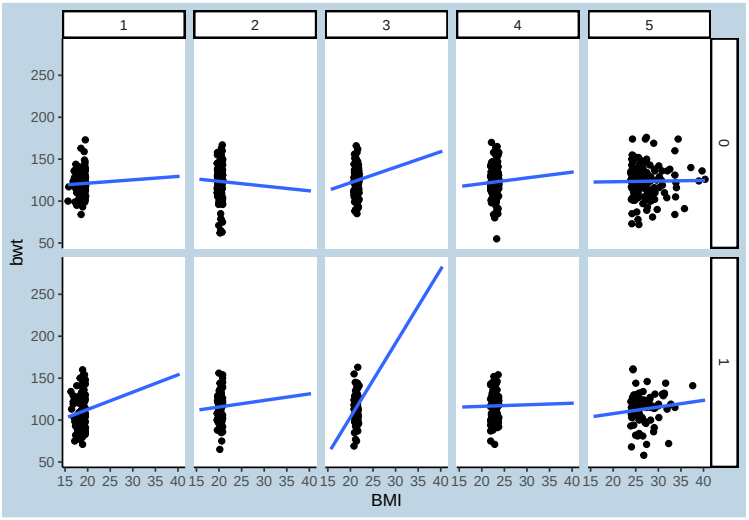


Yes, this graph verify my finding in the previous part.

(k) Does it appear that BMI is related to the birth weight of a baby? Consider making a scatter plot of birth weight vs BMI while conditioning on BMI quartiles and whether the mother smokes to help answer this question. Q25: 2/2

Answer:

```
ggplot(data = cleannp_i, aes(x = BMI, y = bwt)) +
  geom_point() +
  facet_grid(rows = vars(smoke),
             cols = vars(quartile)) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```



No, it appears that BMI is not directly related to the birth weight of a baby.

2. Verify the results of theoretical question 3, parts (c) and (d), by simulating x1, . . . , x25 in R and performing the relevant computations. Q26: 1/2

Answer:

```
x = runif(25, min=0, max=100)
mean_x = mean(x)
y = x - mean_x
sd(y)
```

```
## [1] 32.97288
```

```
sd(x)
```

```
## [1] 32.97288
```

```
var(y)
```

```
## [1] 1087.21
```

```
var(x)
```

```
## [1] 1087.21
```

parts (c) proved.

```
z = y / sd(x)
sd(z)
```

```
## [1] 1
```

```
var(z)
```

```
## [1] 1
```

parts (d) proved.